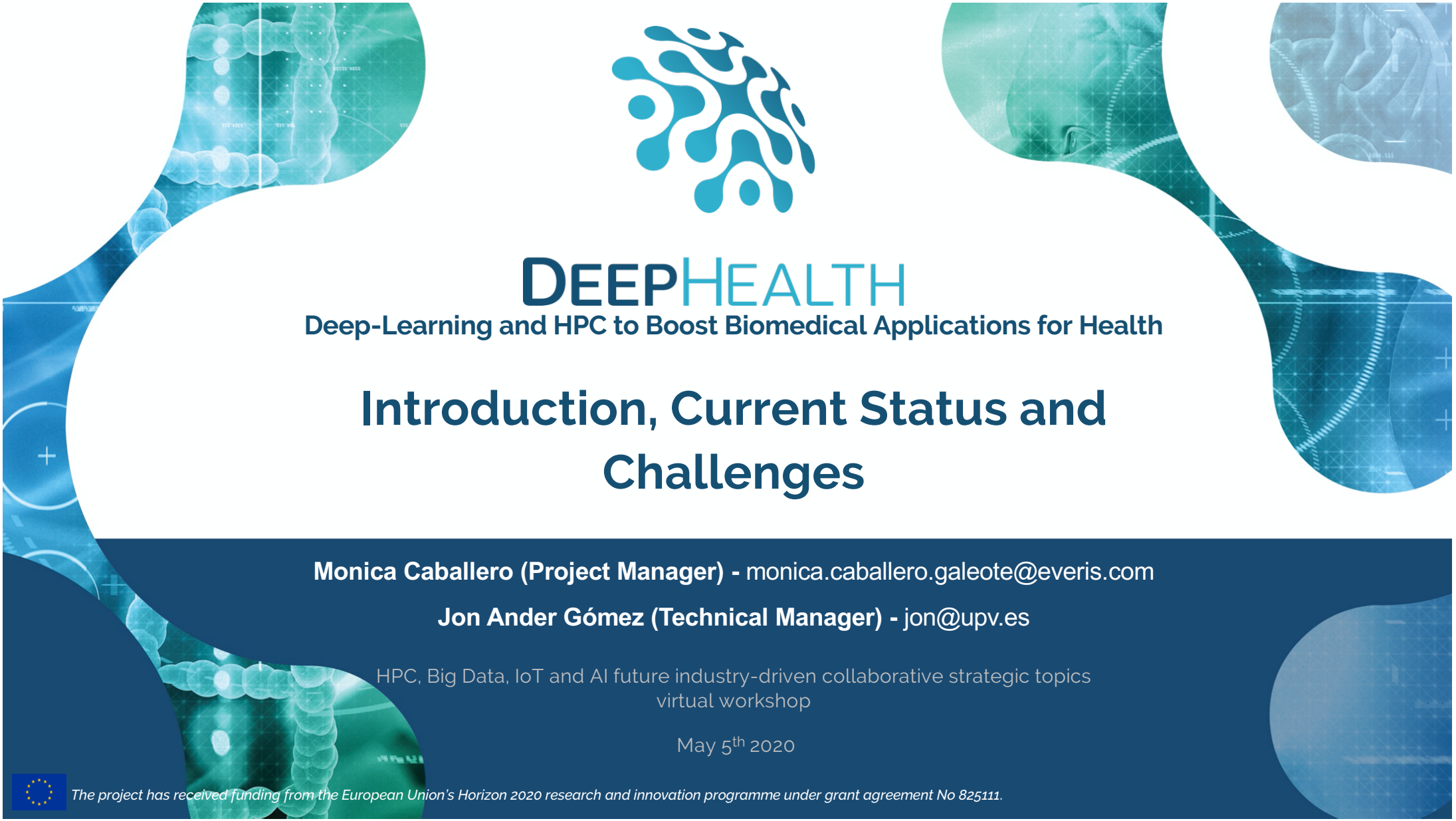


DEEPHEALTH

Deep-Learning and HPC to Boost Biomedical Applications for Health

Introduction, Current Status and Challenges

Monica Caballero (Project Manager) - monica.caballero.galeote@everis.com

Jon Ander Gómez (Technical Manager) - jon@upv.es

HPC, Big Data, IoT and AI future industry-driven collaborative strategic topics virtual workshop

May 5th 2020

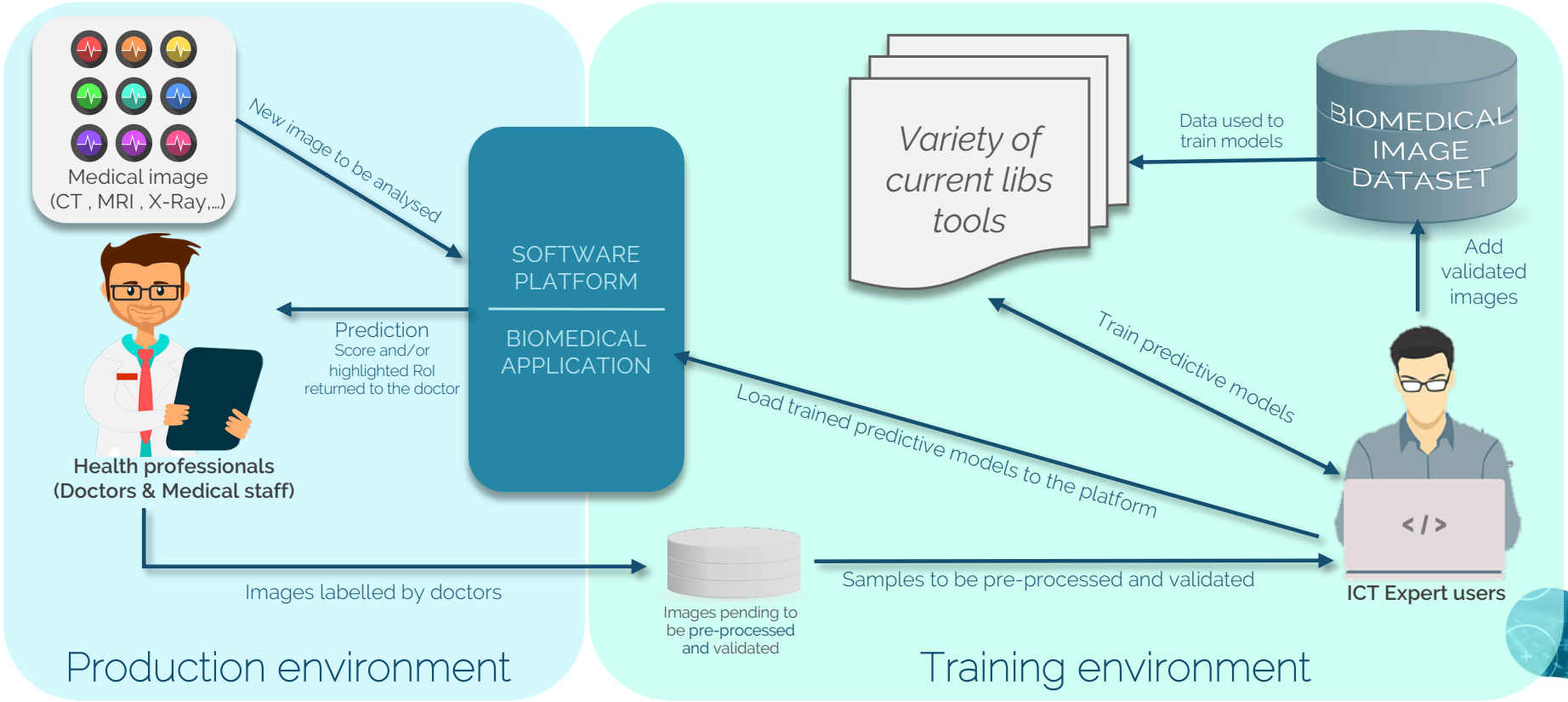
The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

A Little bit of context

- Healthcare: key sector in the global economy
- Public health systems generate large datasets of biomedical images
  - Large unexploited knowledge database
  - Interpretation of the clinical expert manually
- R & D on applying Artificial Intelligence (AI) to analyze biomedical images but . . .
  - Need for advanced skills in AI and different technologies and tools
  - Expensive processes in time and resources
  - Needs of high-quality data and take care of ethics
- HPC and BigData technologies (Big Data, HPC) sufficiently mature and available.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

The scenario



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

About DeepHealth

DEEPHEALTH

Aim & Goals

- **Put HPC computing power at the service of biomedical applications** with DL needs and apply DL techniques on large and complex image biomedical datasets to support **new and more efficient ways of diagnosis, monitoring and treatment of diseases.**
- Facilitate the daily work and **increase the productivity of medical personnel and IT professionals** in terms of image processing and the **use and training of predictive models** without the need of combining numerous tools.
- Offer a **unified framework** adapted to exploit underlying heterogeneous **HPC and Cloud architectures** supporting state-of-the-art and next-generation **Deep Learning (AI) and Computer Vision algorithms** to enhance European-based medical software platforms.

Key facts

Duration: 36 months
Starting date: Jan 2019

Budget 14.642.366 €
EU funding 12.774.824 €

22 partners from **9 countries**:
Research centers, Health organizations, large industries and SMEs

Research Organisations

Universitat Politècnica de València · Barcelona Supercomputing Center · list · UNIMORE · Karolinska Institutet · CRS4 · Università degli Studi di Torino · Otto von Guericke Universität Magdeburg

Health Organisations

Azienda Ospedaliero-Universitaria Città della Salute e della Scienza di Torino · Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana · CHUV · Spitalul Clinic Prof. Dr. Theodor Burghele București · Stockholms läns landsting

Large Industries

everis an NTT DATA Company · PHILIPS · THALES · SIMAVI Software Imagination & Vision

SMEs

WINGS ICT Solutions · TREE Technology · prodesign electronic gmbh · security Stelar research

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

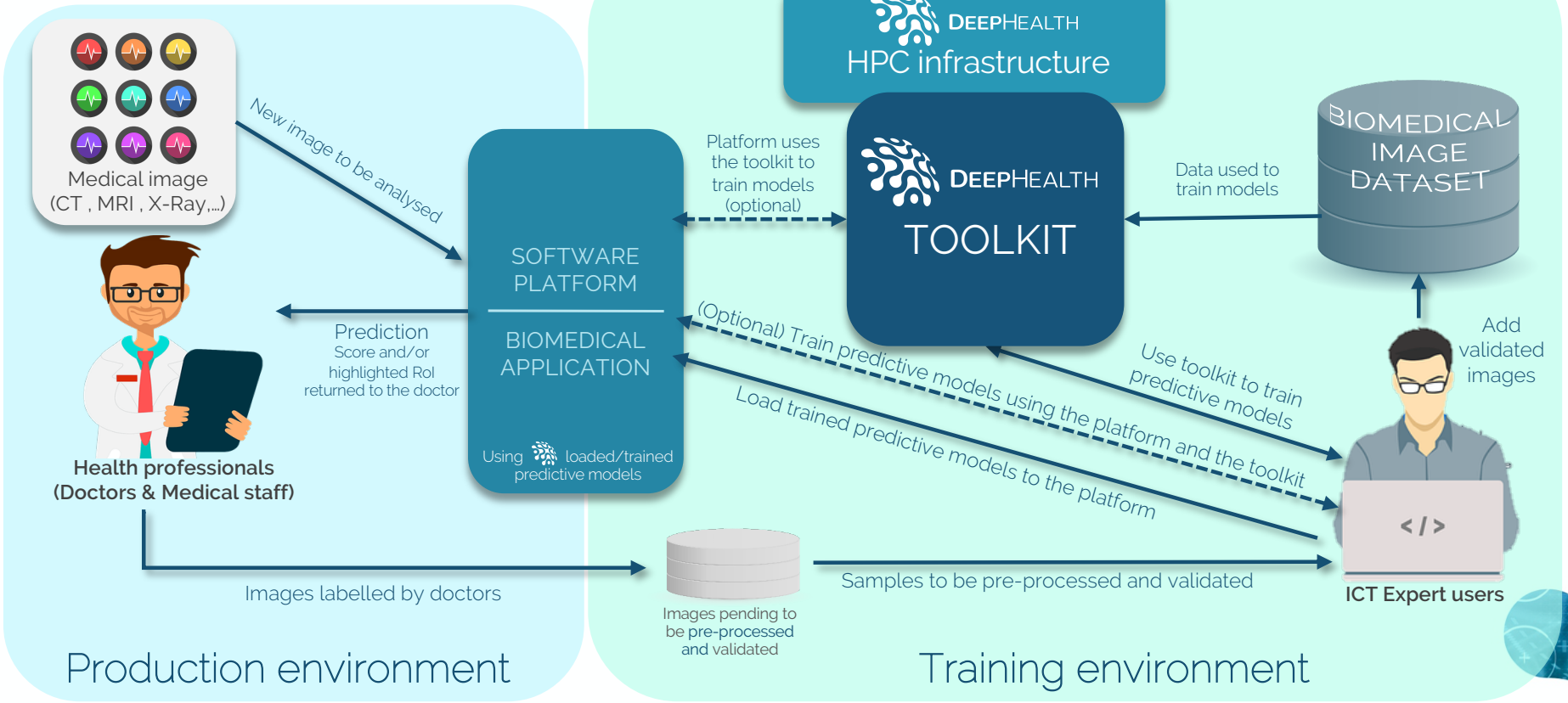
Developments & Expected Results

- **The DeepHealth toolkit**
  - Free and open-source software: 2 libraries + front-end.
    - ***EDDLL:*** *The European Distributed Deep Learning Library*
    - ***ECVL:*** *the European Computer Vision Library*

  - Ready to run algorithms on Hybrid HPC + Cloud architectures with heterogeneous hardware (Distributed versions of the training algorithms)
  - Ready to be integrated into end-user software platforms or applications
- **HPC infrastructure** for an efficient execution of the training algorithms which are computational intensive by making use of heterogeneous hardware in a transparent way
- Seven enhanced **biomedical and AI software platforms** provided by EVERIS, PHILIPS, THALES, UNITO, WINGS, CRS4 and CEA that integrate the DeepHealth libraries to improve their potential
- Proposal for a structure for anonymised and pseudonymised data lakes
- **Validation** in 14 use cases (*Neurological diseases, Tumor detection and early cancer prediction, Digital pathology and automated image annotation*).

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

The Concept



The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Current status at May 2020 (M1–M16)

- **Design & specification** of libraries APIs, HPC system and use-cases finished.
- Definition of the **structure** for anonymised and pseudonymised **data lakes** in progress
- **DeepHealth Toolkit**:
  - **EDDL & ECVL:**
    - First stable versions publicly available in GitHub (non-distributed): https://github.com/deephealthproject/eddl / | https://github.com/deephealthproject/ecvl
    - *Work in progress:* improving non-distributed version, Python wrapper, distributed version using COMPSs from BSC to run on hybrid HPC + Cloud architectures, also using FPGAs kernels
  - **Front-end:**
    - GUI mostly finished
    - Supporting back-end to load and transform images on the fly during training ready

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Current status at May 2020 (M1–M16)

- **HPC Infrastructure Adaptation** – in progress:
  - Development of heterogeneous support for the DeepHealth libraries
  - Communication and Synchronisation optimizations to make training and inference algorithms efficient
  - HPC runtime support and custom hardware support for both libraries
  - Design an hybrid cloud computing solution to support DeepHealth libraries
- Integration of libraries and use cases in **application platforms**— in progress:
  - Adaptation of application platforms to use cases specifics
  - Design of the "how to" integrate libraries into the platforms finished. Iterative integration in progress
- **Testing and pilots**:
  - About to start 1st technical validation setting complete pipelines
    - Example: https://github.com/deephealthproject/use_case_pipeline
  - Use-cases preparing data-sets for pilots (acquiring, labelling, curation, etc.)

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenges identified

1. Need of faster communication networks
2. Availability of Quality & Interoperable Data
3. Data Security & Anonymization
4. Automated end-to-end AI & ML analytic pipelines
5. Domain Application Experts in the loop

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 1: Need of faster communication networks

- **Key topic**: distributed training algorithms need to frequently move model parameters between master node(s) and worker nodes.
- **State of the art and current limitations**: ***InfiniBand*** and other solutions provide low latency communications within HPC environments, but AI/ML/DL training algorithms need to **move large amounts of data too frequently.**
- **Future challenges**: ***research*** — algorithmic strategies to make more asynchronous the distributed training in order to diminish data transfer requirements while convergence is guaranteed, ***innovation*** — new hardware to allow higher bit rate communications.
- **Future opportunities**: development of faster communication hardware and research on distributed training algorithms.
- **Impact**: reduction of distributed training time on HPC environments.
- **Concrete actions**: HW manufacturers to design faster communications hardware (short and long term), DL experts to design and implement new training algorithms (short and long term).

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 2: Availability of Quality & Interoperable Data

- **Key topic**:
  - Data **correctly labelled**, complete, and without any bias is key for training AI/ML/DL models and its **application to clinical purposes** (diagnosis, monitoring and treatment of patients).
  - The **availability of large datasets** related to the same disease coming from different sources (e.g. hospitals of European public health systems), homogeneously annotated and validated, is crucial to the progress of Europe in the application of AI-based solutions to the Health sector.
  - Health data presents also a great variability in terms of formats, structure, etc. ***Interoperability*** is needed for the aggregation and joint use-exploitation of all data available.
- **State of the art and current limitations**:
  - ***Data sharing and interoperability***: There already exist many datasets from the health sector publicly available that are **not compatible among them**, so that those which belong to the same disease or are addressing the same aspect can not be used altogether for training AI-based models. Standard formats being proposed, but an increasing amount of non-structure data is being acquired.
  - ***Data quality***: Large quantity of data not (properly) labelled because doing it is a time consuming task.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 2: Availability of Quality & Interoperable Data

- **Future challenges**:
  - **GDPR limitations for sharing pseudonymised datasets.** With no risks no scientific and technological advances are possible. Law must be adapted to regulate data sharing within the Health sector **to make doctors feel confident** when participating in projects from which **pseudonymised datasets will be made publicly available for research** in addition to train and validate commercial AI-based solutions.
  - **AI applied to ease data labelling and check data correctness**. Semi-automatic labelling systems can help to increasingly make available data sets with quality data.
  - Automate the **structuration and management** of the increasing amount of non-structured data that can be used for Health purposes. Data coming from wearables, social-media, IoT, etc.
- **Future opportunities**: projects focused on both:
  - Definition of guidelines and protocols to collect data sets of quality data.
  - Collecting data sets collaboratively from different hospitals across Europe.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 2: Availability of Quality & Interoperable Data

- **Impact:**
  - Data is key for training **accurate enough AI models**. Different datasets from the same disease properly annotated and compatible in terms of the type of images and how images were acquired can be used to form larger datasets. **Larger datasets** of quality health data (images and other clinical data) **will lead to more robust AI/ML models** to support doctors during diagnosis.
  - **Data sharing** can also impact as an enabler for the generation of new knowledge allowing comparison studies and the use of AI for that (new AI applications).
- **Concrete actions:**
  - ***Short-term:*** Law need to be modified or amended accordin to the needs of scientific and technological progress. **Protocols for acquiring new data must be defined** and shared by all health practitioners in order to make future datasets are compatible to be merged in order to form larger datasets to be shared by the scientific communities, both Health community and AI/ML community.
  - ***Medium-term:*** **innovation actions** must be carried out to promote in hospitals the acquisition of data following standard guidelines. Doctors and other clinical personnel need to be instructed and guided during the transition.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 3: Data Security & Anonymization

- **Key topic**: Health data is a **very sensitive data** that may include patients information. In order to use health data for AI purposes it must be ensured:
  - Data is used appropriately and only by authorized personnel.
  - No patient information is disclosed.
  - Compliance with EU and local regulations of member states.
- **State of the art and current limitations**: Anonymisation, pseudonymisation and de-identification techniques are being developed, as well as techniques such as federated or split learning to ease the use of AI ensuring security and privacy of sensitive data.
- **Future challenges**:
  - Anonymisation techniques future-proof that ensure no future algorithms can identify the patient (too challenging).
  - New training strategies to handle sensitive data and edge-computing ensuring no data disclosures can occur while data is used for training AI-based models.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 3: Data Security & Anonymization

- **Future opportunities**:
  - To spread out data sharing between research institutions and industrial stakeholders.
  - Reach true exascale computing in the Health sector with huge datasets to train more robust AI that can be shared by all the stakeholders in this sector.
- **Impact**: EU builds ethical AI, increase data sharing possibilities, increase trust, etc.
- **Concrete actions**:
  - ***Short-term:*** during the transition: committees must be created to issue authorisations for sharing data sets in order to avoid responsibility on single doctors or teams from a given hospital.
  - ***Medium-term:*** Define guidelines and protocols to collect data taking into account privacy and security by design. Specific H2020 calls can be defined for this purpose.
  - ***Long-term:*** GDPR and other Law regulations should be improved/updated (or new regulations created) so the responsibility of sharing data does not relay on doctors (or teams of doctors) who collected and hold data.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 4: Automated end-to-end AI & ML analytic pipelines

- **Key topic**:
  - Computer engineers/scientists working in the Health sector do not need/might not have a deep knowledge of the AI/ML/DL techniques for their daily work.
  - Instead, they could rely on intuitive software platforms/services (e.g. with a web-based GUI) to carry out all the tasks they have to do in order to train AI/ML/DL models and put trained and tested models in production, i.e. integrating such models in the software platform/application used by doctors.
- **State of the art and current limitations**:
  - H2020 ICT-11 DeepHealth project (GA No 825111) is focused on automating end-to-end AI/ML analytic pipelines, both for training AI/ML models and for using the trained models in real scenarios.
  - Limitations: acceptance of AI-based models to be included in the protocols followed by doctors.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 4: Automated end-to-end AI & ML analytic pipelines

- **Future challenges**:
  - Keep EU competitiveness in front of other toolkits (deal with aspects as compatibility with evolving technologies, future-proof to integrate new algorithms, etc.).
  - Extend analytic pipelines to cover more specific / narrow specifications in health and other domains.
  - Increase acceptance (trust!) in professionals (e.g. clinical professionals).
- **Future opportunities**:
  - Increased number of AI-based applications providing more powerful/accurate services to support health and other professionals.
- **Impact**: Increase of the productivity of computer scientists working in the Health sector. These technical people are becoming more and more essential in the value chain of **smart data use** in the Health sector.
- **Concrete actions**: EU calls addressing the further development of this kind of systems for the health sector and beyond.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 5: Domain Application Experts in the loop

- **Key topic**: Domain application experts are key for developing AI models and AI-based solutions:
  - Collecting quality data needs the effort of domain application experts, people usually too busy to participate in additional tasks or to change their current way of working —This is especially evident in the healthcare sector, where physicians are not always willing to spend time annotating data or avoiding to include patient identification data into images or text.
  - Usefulness and trust in the AI-based application depend on the real involvement of the experts in the design phase: what information/support do they really need? What other information is not to useful and can lead to doubts on the results obtained (e.g. since the difficulty on the task, the lack of explainable results?).
- **State of the art and current limitations**: Experts are involved in projects. While efforts are done for user-centred design approaches, they usually end-up acting as end-users sometimes disconnected and mostly present at the end of the development phase.
  - They tend to be people usually too busy to participate in additional tasks or to change their current way of working.
  - Dialogue between expert domains and technical profiles is still complex.
  - They are attracted by the technology capabilities but there are difficulties to understand the real possibilities so to have more valuable interactions in the early stages of development.
  - Requirements gathering and design phase is still lead most of the times by technical profiles.

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

Challenge 5: Domain Application Experts in the loop

- **Future challenges**:
  - Data annotation, quality: Study current professional workflows to design annotation procedures, methodologies and tools that suit better application experts.
  - Increase interdisciplinary, user-centred development methodologies to bring domain and ICT experts closer.
  - Find common ground and language between domain and ICT experts, intensify connections.
  - Increase the relevance of domain experts in ICT projects workplans
- **Future opportunities**:
  - More powerful projects with real interdisciplinary teams
- **Impact**:
  - Availability of quality data.
  - AI models and AI-based solutions matching real needs, domain expectations and useful to be scaled-up and ported to the market effortlessly.
- **Concrete actions**:
  - Request more and more involvement of domain application experts in ICT projects in future calls.
  - Balance ICT and domain-specific aspects in project calls to attract domain experts

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.

DEEPHEALTH

Questions?

Project Coordinator: Mónica Caballero
monica.caballero.galeote@everis.com

Technical Manager: Jon Ander Gómez
jon@upv.es

https://deephealth-project.eu

@DeepHealthEU

The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825111.